título

Procesamiento masivo de datos mediante Cassandra y Spark

Máster en Sistemas Informáticos Avanzados
Septiembre de 2016

Autor:

Xabier Zabala Barandiaran

Supervisores:

German Rigau i Claramunt
UPV/EHU
Iñigo Etxabe y Beñat Aranburu
Datik Información Inteligente S.L.

Agradecimientos

En primer lugar, quisiera expresar mi gratitud hacia las personas que han posibilitado la concepción y el desarrollo de este proyecto. Agradezco a German Rigau i Claramunt, supervisor del proyecto por parte de la UPV/EHU, la predisposición mostrada y el asesoramiento ofrecido durante el transcurso del mismo. Doy también las gracias a Inigo Etxabe y Beñat Aranburu, supervisores del proyecto por parte de Datik Información Inteligente S.L., por poner a mi disposición todos los medios tecnológicos necesarios para llevar a cabo el proyecto y por el trato ofrecido desde el primer día.

Agradecerles, cómo no, a mis padres José Javier Zabala y María Pilar Barandiaran por el esfuerzo desempeñado para facilitarme, en la medida que les ha sido posible, el camino que he recorrido hasta llegar aquí. Me congratula haber sabido responder satisfactoriamente a la confianza que ellos siempre han depositado en mí. Por todo lo que han hecho y por todo lo que suponen para mí, un beso enorme para los dos.

No quisiera olvidarme de todas las personas que han estado a mi lado durante este maravilloso periplo, a los cuales no me atrevo a mencionar de forma individual por el miedo de dejar a alguno en el tintero. Gracias a los amigos de toda la vida por el apoyo ofrecido durante este camino. Gracias a los compañeros de la facultad, por todo los momentos vividos juntos y en especial a aquellos que durante esta etapa se han ganado a pulso el privilegio a ser parte importante de lo que me resta de existencia.

Por último, pero no por ello menos importante, quisiera evocar a todos los docentes que han tomado parte en mi formación desde aquel Septiembre del 2008 y agradecer a todos ellos el conocimiento compartido y el esfuerzo invertido en mí durante estos años.

Gracias de todo corazón a la gente mencionada en este breve capítulo por haber hecho de mí un mejor profesional y sobre todo una mejor persona.

Resumen

Proyecto Final del Máster en Sistemas Informáticos Avanzados. Estudio comparativo de carácter empírico realizado sobre el rendimiento ofrecido por varias tecnologías emergentes en el área del Big Data a la hora de operar en escenarios que requieren un almacenamiento y procesamiento eficaz de volúmenes masivos de datos.

Se ha construido un clúster compuesto por tres nodos virtuales que operan sobre Linux. Una vez configurados y dotados de tecnología necesaria para el correcto funcionamiento de MySQL Cluster [**mysqlcluster**] y Apache Cassandra [**apachecassandra**], se ha procedido a poblar dichos sistemas de almacenamiento utilizando un data-set público de aproximadamente 25GB. Un conjunto de consultas diseñadas atendiendo a la naturaleza de los datos almacenados han sido ejecutadas sobre ambas infraestructuras. El uso de Apache Spark [**apachespark**] durante el proceso ha posibilitado la distribución de la carga computacional entre los nodos que componen el clúster. A su vez, se ha erigido un nodo virtual con especificaciones técnicas equivalentes a la suma de los 3 nodos virtuales que forman el clúster para emular la respuesta ofrecida por una base de datos MySQL tradicional en el mismo contexto.

El estudio evidencia que, entre las tecnologías de almacenamiento distribuido comparadas, Apache Cassandra es la alternativa más eficaz a la hora de tratar volúmenes masivos de datos a costa de tener que realizar un análisis y diseño previo de los mismos. No obstante, MySQL Cluster es una opción muy a tener en cuenta, ya que sacrificando la eficacia de forma ligera, posibilita estructurar datos de manera más flexible. Ambas tecnologías demuestran estar mejor capacitadas que el MySQL monolítico tradicional para operar en el contexto analizado.

Palabras Clave: Big Data, MySQL Cluster, Apache Cassandra, Apache Spark, Clústering.

Índice general